

Less is More: Micro-expression Recognition from Video using Apex Frame

Sze-Teng Liong^{*1,2}, John See³, Raphael C.-W. Phan², and KokSheik Wong¹

¹Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

²Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia

³Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia

Abstract Despite recent interest and advances in facial micro-expression research, there is still plenty room for improvement in terms of micro-expression recognition. Conventional feature extraction approaches for micro-expression video consider either the whole video sequence or a part of it, for representation. However, with the high-speed video capture of micro-expressions (100-200 fps), are all frames necessary to provide a sufficiently meaningful representation? Is the luxury of data a bane to accurate recognition? A novel proposition is presented in this paper, whereby we utilize only two images per video: the apex frame and the onset frame. The apex frame of a video contains the highest intensity of expression changes among all frames, while the onset is the perfect choice of a reference frame with neutral expression. A new feature extractor, Bi-Weighted Oriented Optical Flow (Bi-WOOF) is proposed to encode essential expressiveness of the apex frame. We evaluated the proposed method on four micro-expression databases—CASME II, SMIC-HS, SMIC-NIR and SMIC-VIS. Our experiments lend credence to our hypothesis, with our proposed technique achieving a state-of-the-art F1-score recognition performance of 61% and 62% in the high frame rate CASME II and SMIC-HS databases respectively.

Keywords: micro-expressions, emotion, apex, optical flow, optical strain, recognition

1 Introduction

Have you ever thought someone was lying to you, but have no evidence to prove it? Or have you always

found it difficult to interpret one's emotion? Recognizing micro-expressions could help to solve these doubts.

Micro-expression is a very brief and rapid facial emotion that is provoked involuntarily [6], revealing a person's true feelings. Akin to normal facial expression, also known as *macro-expression*, it can be categorized into six basic emotions: happy, fear, sad, surprise, anger and disgust. However, macro-expressions are easily identified in real-time situations with the naked eye as it occurs between 2–3 seconds and can be found over the entire face region. On the other hand, a micro-expression is both *micro* (short duration) and *subtle* (small intensity) [7] in nature. It lasts between 1/5 to 1/25 of a second and usually occurs in only a few parts on the face. These are the main reasons why people are sometimes unable to realize or recognize the genuine emotion shown on a person's face [5, 25]. Hence, the ability to recognize micro-expressions is beneficial in both our mundane lives and also society at large. At a personal level, we can differentiate if someone is telling the truth or lie []. Also, analyzing a person's emotions can help facilitate understanding of our social relationships, while we are increasingly awareness of the emotional states of our own selves and of the people around us. More essentially, recognizing these micro-expressions is useful in a wide range of applications, including psychological and clinical diagnosis, police interrogation and national security [9, 24, 10].

"Micro-expression" was first discovered by psychologists, Ekman and Friesen [6] in 1969, from a case where a patient was trying to conceal his sad feeling by covering up with smile. They detected the patient's genuine feeling by carefully observing the subtle movements on his face, and found out that

^{*}Corresponding author.

Email: szeteng@siswa.um.edu.my

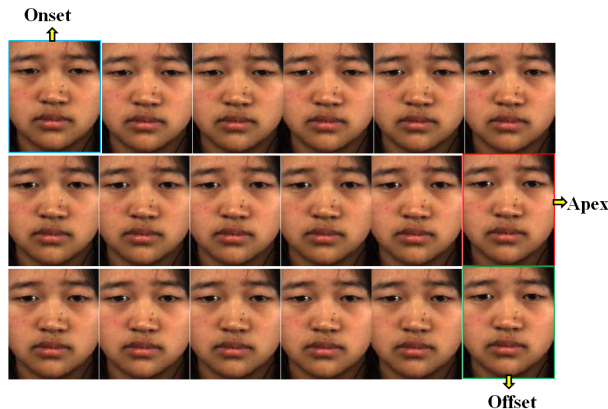


Figure 1: Example of a sequence of image frames (ordered from left to right, top to bottom) of a happiness expression in from the CASME II [33] database, with the onset, apex and offset frame indications

the patient was actually planning to commit suicide. Later on, they established Facial Action Coding System (FACS) [8] to determine the relationship between facial muscle changes and emotional states. This system can be used to identify the exact time each action unit (AU) begins and ends. The occurrence of the first visible AU is called the onset, while that of the disappearance of the AU is the offset. Apex is the point when the AU reaches the peak or the highest intensity of the facial motion. The timings of the onset, offset and apex for the AUs may differ for the same emotion type. Figure 1 shows a sample sequence containing frames of a happiness expression from a micro-expression database, with the indication of onset, apex and offset frames.

Micro-expression analysis is arguably one of the less explored area of research in the field of computer vision. Currently, there are less than fifty micro-expressions related research papers published since 2009. In contrary, the feature extractions method and the databases in macro-expression studies are well-developed and established. Thus, there is a lack of micro-expression database for the technique proposed to be evaluated and analyzed on. It also hinders the progress on micro-expression researches. Moreover, the subtlety, minuteness and the quickness of the

micro-expression occurrence, are also the obstacles encountered in the research process.

There are two types of primary studies on micro-expression system, i.e., spotting and recognition. The former is to indicate the interval of micro-expression occurrence or the frame indices of some important instants (such as onset, apex and offset). The latter is to classify the expression type given a micro-expression video sequence. Majority of the articles focused on the micro-expression recognition system, that mainly implement new feature extraction methods to improve the micro-expression recognition rate.

All the micro-expression databases are pre-processed before being released to the public. This process includes face registration, face alignment and groundtruth labeling (i.e., AU, emotion type, frame indices of onset, apex and offset). In the two most popular spontaneous micro-expression databases, CASME II [33] and SMIC [17], common procedures of face registration and alignment algorithms used were: (1) Active Shape Model (ASM) [2]: a set of landmark coordinates were detected; (2) Local Weighted Mean (LWM) [11]: the faces were transformed based on the template face according to the landmarks points. Note that both ASM and LWM operate fully automatically. However, the last process, groundtruth labeling, is not automated and are done with the help of psychologists or trained person. In other words, the results of groundtruth labeling may be varied depending on the coders. As such, the reliability and consistency of the marking are directly affected. As a consequence, imprecise groundtruth information may influence the recognition accuracy of the micro-expression recognition system.

There are several works in literature which attempted to spot the temporal interval (onset-offset) containing micro-expressions from raw videos in the databases. By *raw*, we refer to video clips in its original captured form, without pre-processing. In [22], the authors searched for the frame indices that contain micro-expressions. They utilized Chi-Squared dissimilarity to calculate the distribution difference between the Local Binary Pattern (LBP) histogram of the current feature frame and the averaged feature frame. The frames which fall above a certain threshold were regarded as micro-expression frames.

A similar approach was carried out by [4], except that a denoising method was added before extracting the features, and that the Histogram of Gradient was used instead of LBP. However, the database they tested on was not publicly available. Since the benchmark used in this paper and in [22] are different, the performance of these two methods cannot be compared directly. Both of these papers claimed that the eye blinking movement is one of the micro-expressions. However, it was not mentioned in the ground-truth details and hence the frames that contain eye blinking was annotated manually.

To the best of our knowledge, there is only one prior work that combines micro-expression spotting and recognition system. It was implemented by Li et al. [16]. They extended the work of [22], whereby after the spotting stage, the spotted micro-expression frames (with the onset and offset information) were concatenated to a single sequence for expression recognition. In the recognition task, they employed motion magnification technique and proposed a new feature extractor - the Histograms of Image Gradient Orientation. However, the recognition performance was poor compared to the state-of-the-art. Besides, the frame rate of the database is 25fps, which means that the maximum frame number in a raw video sequence is only 5.

Apart from the aforementioned micro-expression frames searching approaches, the other technique used is to automatically spot the instance of the single apex frame in a video. The micro-expression details retrieved from that apex frame are expected to be helpful in both psychological and computer vision research purpose, because it contains the maximum facial muscle movements throughout the video sequence. Yan et al. [32] published the first work in spotting the apex frame. They employed two feature extractors (i.e., LBP and Constraint Local Models) and reported the average frame distance between the spotted apex and the ground-truth apex. The frame that has the highest feature difference between the first frame and the subsequent frames is the apex. There are two flaws in this work: (1) The average frame distance calculated was not in absolute mean, and led to incorrect results computation; (2) The method was validated using only $\sim 20\%$ of the video

samples in the databases, hence it was by no means conclusive.

The second work on apex frame spotting was presented by Liong et al. [20]. The main differences between this and the previous work [32] are: (1) A *binary search* strategy was implemented to locate the frame index of the apex, because the maximum difference between the first and the subsequent frames might not necessarily be the apex frame; (2) An extra feature extractor was added to confirm the reliability of the method proposed; (3) Certain important facial regions were considered for feature encoding instead of the whole face; (4) All the video samples in the database were used for evaluation and the average frame distance between the spotted and groundtruth apex reported were in absolute mean.

In this paper, we propose a novel approach to micro-expression recognition, whereby for each video sequence, we encode features from the representative apex frame with the onset frame as the reference frame. The onset frame is assumed to be the neutral face and is provided in all micro-expression databases (CASME II and SMIC) while the apex frame labels are only available in CASME II. To solve the lack of apex information in the SMIC, a binary search strategy was employed to spot the apex frame [20]. In this paper, we renamed *binary search* to *divide-and-conquer* for a more general terminology to this scheme. Additionally, we introduce a new feature extractor, Bi-Weighted Oriented Optical Flow (Bi-WOOF) which is capable of representing the apex frame in a discriminative manner which emphasizes facial motion information at both bin and block levels. The histogram of optical flow orientations is weighted twice at different representation scales; bins by the magnitudes of optical flow, and block regions by the magnitudes of optical strain. Optical flow estimates the motion of objects between two images over time, which was found effective in a recently proposed micro-expression recognition system [21]. Optical strain, which is an extension of the optical flow, provides more precise subtle and micro facial changes, as was proven useful in several works on micro-expression analysis [27, 26, 18, 19]. We establish our proposition by proving experimentally through a comprehensive evaluation that was carried

out on four notable databases.

The rest of the paper is organized as follows. Section 2 explains the proposed algorithm in detail. The description of the databases used are discussed in Section 3, followed by 4 that reports the experimental results and discussion for the recognition of micro-expressions. Finally, conclusion is summarized in Section 5.

2 Proposed Algorithm

The proposed methodology of the micro-expression recognition system is made up of two components: apex frame spotting and micro-expression recognition. The overall system framework is illustrated in 2. Details of each step will now be elaborated in the following subsections.

2.1 Apex Spotting

For the apex frame spotting task, we employ the apex spotting approach proposed by Liong et al. [20]. The method consists of five steps: (1) The facial landmark points are first annotated using a landmark detector, Discriminative Response Map Fitting (DRMF) [1]; (2) The regions of interest that indicate the facial region with important micro-expression details are extracted according to the landmark coordinates; (3) The LBP feature descriptor is utilized to obtain the features of each frame in the video sequence (i.e., from onset to offset); (4) The feature difference between the onset and the rest of the frames are computed using the correlation coefficient formula; (5) Then a peak detector with *divide-and-conquer* strategy is used to search for the apex frame based on the LBP feature difference. Fig. 3 demonstrates the apex frame spotting approach in a sample video. It can be seen that, the distance of the ground-truth apex (frame #63) and the spotted apex (frame #64) is only difference by one frame.

2.2 Micro-expression Recognition

Here, we discuss a new feature extractor, Bi-Weighted Oriented Optical Flow (Bi-WOOF) that represent each subtle expressions sequence by only

two frames. As illustrated in Fig. 4, the recognition algorithm contains three main steps: (1) The horizontal and vertical optical flow vectors between the apex and neutral frames are estimated; (2) The orientation, magnitude and optical strain of each pixel’s location are computed from the two optical flow components; (3) A Bi-WOOF histogram is formed based on the orientation, with magnitude locally weighted and optical strain globally weighted.

2.2.1 Optical flow estimation

Optical flow approximates the changes of an object’s position between two frames that are sampled at slightly different time. It encodes the motion of an object in vector notation, which indicates the direction and intensity of the flow of each image pixel. The horizontal and vertical components of the optical flow are defined as:

$$\vec{p} = [p = \frac{dx}{dt}, q = \frac{dy}{dt}]^T \quad (1)$$

where dx indicates the changes of the two-dimensional positions and dt is the change in time. The principle of estimating the optical flow is, given a pixel in the first frame, look for the nearby pixel with the same color in the subsequent frame. However, there are three assumptions when estimating the optical flow: (1) Brightness constancy: the observed appearance of an object in the image remains constant over time. Note that, we ignore the apparent motion that are caused by the lighting changes without any actual motion; (2) Spatial coherence: the neighboring pixels with similar pixel intensities in an image are likely to belong to the same region and move in a similar manner; (3) Temporal persistence: the motions of an object between two frames is small and hence the displacements are small. The optical flow constraint equation is given by:

$$\nabla I \bullet \vec{p} + I_t = 0, \quad (2)$$

where $I(x, y, t)$ denotes the temporal image brightness changes in intensity values with respect to time at point (x, y) . $\nabla I = (I_x, I_y)$ are the spatial gradients and I_t is the temporal gradient of the intensity functions.

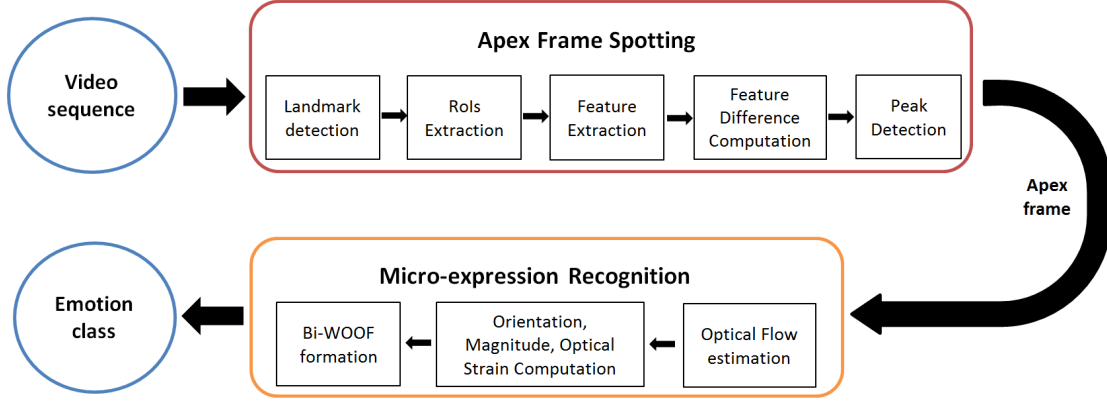


Figure 2: Framework of the proposed micro-expression recognition system

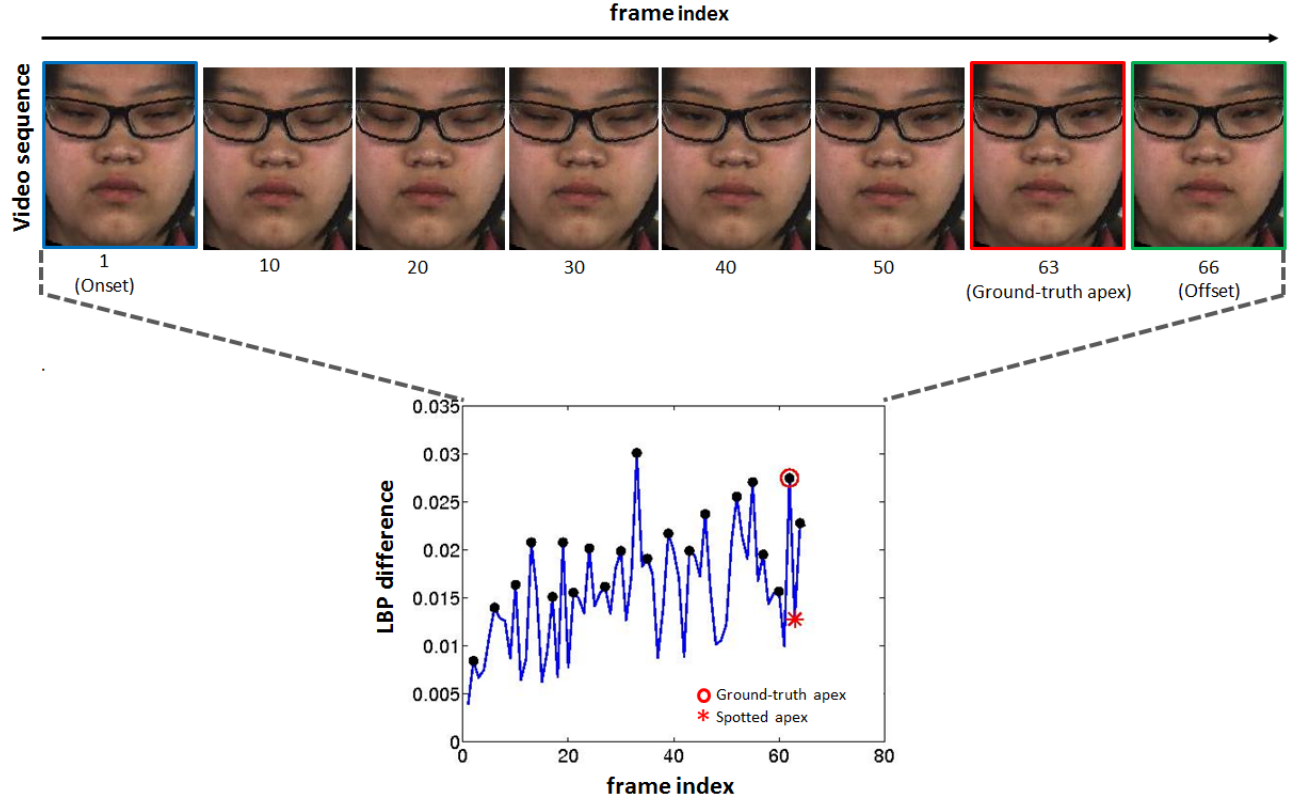


Figure 3: Demonstration of the apex spotting in a video sequence using LBP feature extractor with *divide-and-conquer* strategy

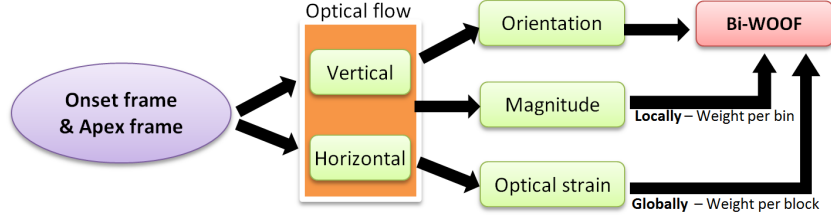


Figure 4: Flow diagram of micro-expression recognition system

We employed TV-L1 [34] for optical flow approximation due to its two major advantages: better noise robustness and the ability to preserve flow discontinuities.

We first introduce and describe the notations that used in the subsequent sections. A micro-expression video clip is denoted as:

$$s_i = \{f_{i,j} | i = 1, \dots, n; j = 1, \dots, F_i\} \quad (3)$$

where F_i is the total number of frames in the i -th sequence, which is taken from a collection of n video sequences. For each video sequence, there is only one apex frame, $f_{i,a} \in f_{i,1}, \dots, f_{i,F_i}$, and it can be situated at any frame index.

The optical flow vectors of the onset (assumed as neutral expression) and the apex frames are predicted, denoted by $\{f_{i,1}, f_{i,a}\}$ respectively. Hence, each video of resolution $X \times Y$ produces only one set of optical flow map, expressed as:

$$\nu_i = \{(u_{x,y}, v_{x,y}) | x = 1, \dots, X; y = 1, \dots, Y\} \quad (4)$$

for $i \in 1, \dots, n$.

2.2.2 Orientation, magnitude and optical strain computation

Given the optical flow vectors, we derive three characteristics to describe the facial motion patterns: (1) magnitude: intensity of the pixel's movement; (2) orientation: direction of the flow motion; (3) optical strain: subtle deformation intensity.

In order to obtain the magnitude and orientation, the flow vectors, $\vec{p} = (p, q)$, are converted from euclidean coordinates to polar coordinates:

$$\rho_{x,y} = \sqrt{p_{x,y}^2 + q_{x,y}^2} \quad (5)$$

$$\theta_{x,y} = \tan^{-1} \frac{q_{x,y}}{p_{x,y}} \quad (6)$$

where ρ and θ are the magnitude and orientation respectively.

The next step is to compute the optical strain, ε , based on the optical flow vectors. For a small enough facial pixel's movement, it is able to approximate the deformation intensity, also known as the infinitesimal strain tensor. In brief, the infinitesimal strain tensor is derived from the Lagrangian and Eulerian strain-tensor after performing a geometric linearisation [28]. In terms of displacements, the typical infinitesimal strain (ε) is defined as:

$$\varepsilon = \frac{1}{2} [\nabla \mathbf{u} + (\nabla \mathbf{u})^T] \quad (7)$$

or can be re-written as:

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2} \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) \\ \varepsilon_{yx} = \frac{1}{2} \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix} \quad (8)$$

where the diagonal strain components, $(\varepsilon_{xx}, \varepsilon_{yy})$, are normal strain components and $(\varepsilon_{xy}, \varepsilon_{yx})$ are shear strain components. Normal strain measures the changes in length along a specific direction, whereas shear strains measures changes in two angles directions that form the plane experiencing shear distortion [31].

To estimate the strain from the optical strain magnitude (Eq. (7)), we can simplify the optical flow vectors (p, q) in (Eq. (2)) by differentiating it to the first order derivatives as the strain components are de-

scribed in a function of displacement vectors (u, v) and shear strain components, expresses as follows: respectively. Specifically,

$$p = \frac{dx}{dt} = \frac{\Delta x}{\Delta t} = \frac{u}{\Delta t}, u = p\Delta t, \quad (9)$$

$$q = \frac{dy}{dt} = \frac{\Delta y}{\Delta t} = \frac{v}{\Delta t}, v = q\Delta t \quad (10)$$

where Δt is the time interval between two image frames. Since the temporal resolution of the a video is constant, Δt is a fixed length, we can approximate the partial derivatives of Eq. (9) and (10) as:

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial p}{\partial x} \Delta t, & \frac{\partial u}{\partial y} &= \frac{\partial p}{\partial y} \Delta t, \\ \frac{\partial v}{\partial x} &= \frac{\partial q}{\partial x} \Delta t, & \frac{\partial v}{\partial y} &= \frac{\partial q}{\partial y} \Delta t \end{aligned} \quad (11)$$

The second order derivatives can be approximated by using finite Difference Approximation.

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{u(x + \Delta x) - u(x - \Delta x)}{2\Delta x} \\ &= \frac{p(x + \Delta x) - p(x - \Delta x)}{2\Delta x} \\ \frac{\partial v}{\partial y} &= \frac{v(y + \Delta y) - v(y - \Delta y)}{2\Delta y} \\ &= \frac{q(y + \Delta y) - q(y - \Delta y)}{2\Delta y} \\ \frac{\partial u}{\partial y} &= \frac{u(y + \Delta y) - u(y - \Delta y)}{2\Delta y} \\ &= \frac{p(y + \Delta y) - p(y - \Delta y)}{2\Delta y} \\ \frac{\partial v}{\partial x} &= \frac{v(x + \Delta x) - v(x - \Delta x)}{2\Delta x} \\ &= \frac{q(x + \Delta x) - q(x - \Delta x)}{2\Delta x} \end{aligned} \quad (12)$$

where $(\Delta x, \Delta y)$ are preset distances of one pixel.

The optical strain magnitude for each pixel can be calculated by taking the sum of squares of the normal

$$\begin{aligned} |\varepsilon_{x,y}| &= \sqrt{\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + \varepsilon_{xy}^2 + \varepsilon_{yx}^2} \\ &= \sqrt{\frac{\partial u^2}{\partial x} + \frac{\partial v^2}{\partial y} + \frac{1}{2} \left(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial x} \right)^2} \end{aligned} \quad (13)$$

2.2.3 Bi-Weighted Oriented Optical Flow

In this stage, we utilize the three aforementioned characteristics (i.e., orientation, magnitude and optical strain images for every video) to build a block-based Bi-Weighted Oriented Optical Flow.

The three characteristic images are partitioned equally into $N \times N$ non-overlapping blocks. For each block, the orientations, $\theta_{x,y}$, are binned and locally weighted according to its magnitude $\rho_{x,y}$. The values of the possible vector orientations fall between $\theta_{x,y} = [-\pi, \pi]$. Thus, the range of each histogram bin is:

$$-\pi + \frac{2\pi c}{C} \leq \theta_{x,y} < -\pi + \frac{2\pi(c+1)}{C} \quad (14)$$

where bin $c \in 1, \dots, C$, and C denotes the total number of histogram bins.

To obtain the global weight ζ_{b_1, b_2} for each block, we utilize the optical strain magnitude, $\varepsilon_{x,y}$:

$$\zeta_{b_1, b_2} = \frac{1}{HL} \sum_{y=(b_2-1)H+1}^{b_2H} \sum_{x=(b_1-1)L+1}^{b_1L} \varepsilon_{x,y} \quad (15)$$

where $L = \frac{X}{N}$, $H = \frac{Y}{N}$, the block indices $(b_1, b_2) \in 1 \dots N$, and (X, Y) are the dimensions (width and height) of the image.

Lastly, the coefficients of ζ_{b_1, b_2} are multiplied with the locally weighted histogram bins to their corresponding blocks. The histogram bins of each block are concatenated to form the resultant feature histogram.

Different from the conventional Histogram of Oriented Optical Flow (HOOF) [3], the orientation histogram bins have equal votes. Here, we consider both the magnitude and optical strain values as the weighting schemes to highlight the importance of each optical flow. Hence, a larger intensity of the pixel's

movement or deformation contributes more effect on the histogram, whereas noisy optical flows with small intensities suppress the significance of the features.

The whole process flow of obtaining the locally and globally weighted features is graphically shown in Fig. 5.

3 Experiment

3.1 Datasets

To evaluate the performance of the proposed algorithm, the experiments were carried out on four spontaneous micro-expression databases, namely CASME II [33], SMIC-HS [17], SMIC-VIS [17] and SMIC-NIR [17]. Note that all these databases are recorded in a constrained laboratory condition due to the subtlety of the micro-expressions.

3.1.1 CASME II

CASME II consists of five kinds of expressions: surprise (25 samples), repression (27 samples), happiness (23 samples), disgust (63 samples) and others (99 samples). Each video clip contains only one micro-expression. Thus, there is a total of 246 video sequences. The emotion labels were marked by two coders with the reliability of 0.85. The expressions were elicited from 26 subjects with the mean age of 22 years old. The camera they used to record the videos was Point Grey GRAS-03K2C. The image resolution and frame rate of the camera were 640×480 and $200fps$ respectively. This database provides the cropped video sequences, meaning only the face exists while the unnecessary background has been eliminated. The cropped images have an average spatial resolution of 170×140 pixels. The average frames per video is 68 (0.34s). The video with the highest frame number is 141 (0.71s) while the lowest is 24 frames (0.12s). The index frames of onset, apex and offset for each video sequence are provided. To perform the recognition task on this micro-expression dataset, the feature extractor employed was block-based LBP-TOP. Then the features were classified by a Support Vector Machine (SVM) with leave-one-video-out cross-validation (LOVOCV) protocol.

3.1.2 SMIC

SMIC includes three sub-datasets, which are SMIC-HS, SMIC-VIS and SMIC-NIR. The data composition of the three datasets are tabulated in detail in Table 1. All the eight participants appeared in VIS and NIR datasets were also involved in HS dataset elicitation. During the recording process, the three cameras (i.e., HS, VIS and NIR) were in used at the same time. The cameras were placed parallel to each other at the middle top of the monitor. The groundtruth of the frame indices of onset and offset for each video clip in SMIC are given, but not apex frame. The three-class recognition task was carried out in the three SMIC datasets individually by utilizing block-based LBP-TOP as the feature extractor and SVM-LOSOCV (leave-one-subject-out cross-validation) as the classifier.

3.1.3 Experiment Settings

In this section, we present the details of the measurement protocol of the experiments, specifically the evaluation method and the parameter settings in the proposed method.

The aforementioned databases (i.e., CASME II and SMIC) are having imbalance distribution of the emotion types. Therefore, it is necessary to measure the recognition performance of the proposed method using F-measure, also suggested in [14]. The equation of F-measure is defined as follows:

$$F\text{-measure} := 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

$$\text{Recall} := \frac{TP}{TP + FN} \quad (17)$$

$$\text{Precision} := \frac{TP}{TP + FP} \quad (18)$$

where TP, FN and FP are true positive, false negative and false positive, respectively.

On the other hand, to avoid person dependent issue in the classification process, we employed LOSOCV strategy in the linear SVM classifier setting. In LOSOCV, the features of the sample videos in one

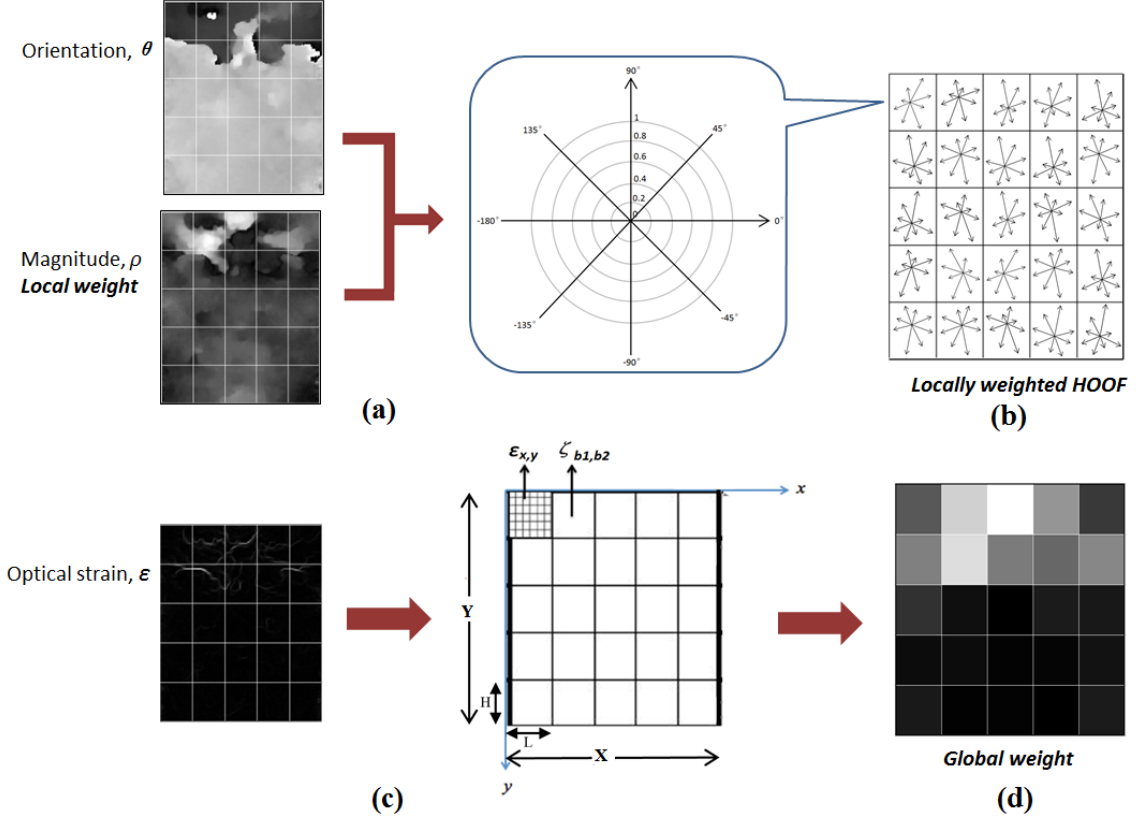


Figure 5: The process of Bi-WOOF feature extraction for a video sample: (a) θ and ρ images are divided into $N \times N$ blocks. In each block, the values of ρ for each pixel are treated as local weights to multiply with their respective θ histogram bins; (b) It forms a locally weighted HOOF with feature size of $N \times N \times C$; (c) $\zeta_{b1,b2}$ denotes the global weighting matrix, which is derived from ϵ image; (d) Finally, $\zeta_{b1,b2}$ are multiplied with their corresponding locally weighted HOOF.

subject are treated as the testing data and the remaining images as the training data. Then, this process is repeated for k times, where k is the number of subjects in the database. Finally, the recognition results for all the subjects are averaged to produce the resultant recognition rate.

For the block-based feature extraction methods (i.e., LBP, LBP-TOP and proposed algorithm), we standardized the block sizes to 5×5 and 8×8 for SMIC and CASME II respectively, as we discovered that these block settings generated reasonably good recognition performance in all cases.

4 Results and Discussion

4.1 Results

The micro-expression recognition performance of the proposed method (i.e., Bi-WOOF) and the other recent feature extraction methods are shown in Table 2. Note that methods #1 to #11 considered all the images in the video sequence (i.e., frames from onset to offset). However for method #12 to #15, only two images were taken to extract the features (i.e., the apex and the onset frames).

To further confirm the importance of the apex

Table 1: Detailed information of the SMIC-HS, SMIC-VIS and SMIC-NIR datasets

		SMIC-HS	SMIC-VIS	SMIC-NIR
Participants		16	8	8
Camera	Type	PixeLINK PL-B774U	Visual camera	Near-infrared camera
	Frame rate (<i>fps</i>)	100	25	25
Expression	Positive	51	28	28
	Negative	70	23	23
	Surprise	43	20	20
	Total	164	71	71
Image resolution	Raw	640×480	640×480	640×480
	Cropped (average)	170×140	170×140	170×140
Frame number	Average	34	10	10
	Maximum	58	13	13
	Minimum	11	4	4
Video duration (<i>s</i>)	Average	0.34	0.4	0.4
	Maximum	0.58	0.52	0.52
	Minimum	0.11	0.16	0.16

frame, we randomly selected one frame in each video sequence and computed the features between that frame and onset using both LBP and Bi-WOOF. The recognition performances of this random frame selection methodology are shown in methods #12 and #14. This process was repeated for 10 times. We observe that utilizing the apex frame is always better than the random ones. As such, it can be concluded that the apex frame plays an important role in contributing discriminant features.

For method #10, LBP-TOP, also known as the baseline, we re-constructed the experiments on the four datasets based on the original papers [17, 33]. On the other hand, we employed Bi-WOOF on all the images in the video sequence. The features were computed by first estimating the three characteristics of the optical flow (i.e., orientation, magnitude and optical strain) between the onset and the subsequent frames (i.e., $\{f_{i,1}, f_{i,j}\}, j \in 2, \dots, F_i$). Next, Bi-WOOF was applied on each image in the video and obtained the resultant histogram. The recognition performance is reported in method #11.

For the LBP feature extractor (i.e., methods #12 and #13), we first obtained the difference image by

simply perform the subtraction between the apex/random frame and the onset frame. This operation is to remove the person identity while preserving the characteristics of facial micro-movements. Then LBP was applied on the difference image to compute the features.

In Table 2, it is noticed that the proposed algorithm, #15 achieves promising results in all the four datasets. More concisely, it outperformed all the other methods in CASME II and SMIC-HS. In addition, for SMIC-VIS and SMIC-NIR, the results of the proposed method are as good as #9, Xu et al. method.

4.2 Discussion

4.2.1 Detailed Analysis on the Recognition Performance

The confusion matrices for the recognition performances on the high frame rate databases, CASME II and SMIC-HS are shown in Table 3 and 4 respectively. It can be seen that there are large classification improvements on all kinds of expressions when employing Bi-WOOF (apex & onset), compared to the baselines. More concretely, in CASME II, the recog-

Table 2: Comparison of micro-expression recognition performance in F-measure measurement on the CASME II, SMIC-HS, SMIC-VIS and SMIC-NIR databases of the state-of-the-art feature extraction methods, random frame selection methodology and the proposed algorithm

Methods		CASME II	SMIC-HS	SMIC-VIS	SMIC-NIR
Whole sequence	1 Le et al. [14]	.33	.47	-	-
	2 Le et al. [13]	.51	-	-	-
	3 Le et al. [15]	.51	.60	-	-
	4 Wang et al. [29]	.40	.55	-	-
	5 Liong et al. [18]	-	.45	-	-
	6 Liong et al. [19]	.38	.54	-	-
	7 Oh et al. [23]	.43	.35	-	-
	8 Huang et al. [12]	.57	.58	-	-
	9 Xu et al. [30]	.30	.54	.60	.60
	10 LBP-TOP [17, 33]	.39	.39	.39	.40
	11 Bi-WOOF	.56	.53	.62	.57
2 images (apex & onset)	12 LBP (random & onset)	.38	.40	.48	.51
	13 LBP (apex & onset)	.41	.45	.49	.54
	14 Bi-WOOF (random & onset)	.50	.46	.56	.50
	15 Bi-WOOF (apex & onset)	.61	.62	.58	.58

niton rate of surprise, disgust, repression, happiness and other expressions increased by 44%, 30%, 22%, 13% and 4% respectively. Furthermore, for SMIC-HS, the recognition rate of the expressions: negative, surprise and positive improved by 31%, 19% and 18% respectively.

Figure 6 illustrates the example of the optical flow derived images between onset and apex frames of a video. The micro-expression shown in the figure is surprise. Referring to the labeling criteria of the emotion in [33], the facial muscle changes are centering at the eyebrow regions. We can hardly tell the facial movements in Figure 6a, 6b and 6c. For Figure 6d, noticeable amount of the muscular changes are taken place at the upper part of the face, whereas in figure 6e, the eyebrows regions have obvious facial movement. Magnitude information emphasizes the amplitude of the facial changes and thus we exploit it as local weight. Due to higher derivative of computation is involved in obtaining optical strain magnitudes, it has the ability to remove the noise and preserve large motion changes. We make use

of its characteristics to build the global weight. Besides, [18] demonstrated that optical strain globally weighted on the LBP-TOP features produced better recognition results compared to without the weighting.

From the evaluation measurements of F-measure and confusion matrices, it has been proven that extracting the features of two images only (i.e., apex and onset frame) using the proposed method, Bi-WOOF is able to yield excellent recognition performance in micro-expression databases, especially in CASME II and SMIC-HS, that have high temporal resolution (i.e., $\geq 100fps$).

4.2.2 Spotting the apex frame

The very first and the most essential criteria to employ the proposed method is the acquisition of the apex index for each video sequence. Although the SMIC datasets (i.e., HS, VIS and NIR) did not provide the ground-truth apex frame index, we utilized *divide-and-conquer* strategy [20] to search for the apex frame. As such, the lack of the apex frame

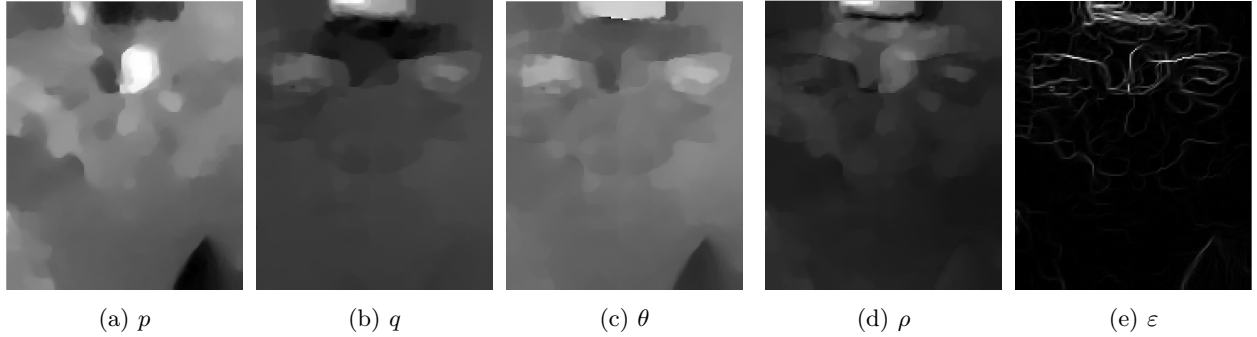


Figure 6: Illustration of the optical flow derived components between onset and apex frames of a video: (a) Horizontal vector of optical flow, p ; (b) Vertical vector of optical flow, q ; (c) Orientation, θ ; (d) Magnitude, ρ ; (e) Optical strain, ε

Table 3: Confusion matrices of baseline and Bi-WOOF (apex & onset) for recognition task on CASME II database, where the emotion types are, DIS: disgust; HAP: happiness; OTH: others; SUR: surprise; REP:repression

(a) Baseline					
	DIS	HAP	OTH	SUR	REP
DIS	.20	.11	.66	.02	.02
HAP	.09	.47	.25	0	.19
OTH	.21	.12	.58	.08	0
SUR	.12	.36	.20	.32	0
REP	.07	.33	.26	.04	.30

(b) Bi-WOOF (apex & onset)					
	DIS	HAP	OTH	SUR	REP
DIS	.49	.07	.44	0	0
HAP	.03	.59	.28	.03	.06
OTH	.21	.09	.62	.01	.06
SUR	.04	.12	.08	.76	0
REP	.07	.19	.22	0	.52

Table 4: Confusion matrices of baseline and Bi-WOOF (apex & onset) for recognition task on SMIC-HS database, where the emotion types are, NEG: negative; POS: positive; SUR:surprise

(a) Baseline			
	NEG	POS	SUR
NEG	.34	.29	.37
POS	.41	.39	.20
SUR	.37	.19	.44

(b) Bi-WOOF (apex & onset)			
	NEG	POS	SUR
NEG	.66	.23	.11
POS	.27	.57	.16
SUR	.23	.14	.63

4.2.3 Computational Cost

Moreover, we examine the computational efficiency of Bi-WOOF in SMIC-HS database on both the *whole sequence* and *two images* (i.e., *apex* and *onset*), which are the methods #11 and #15 in Table 2 respectively. The average duration taken per video of executing the micro-expression recognition system for the *whole sequence* and *two images* in MATLAB implementation

information issue had been resolved.

were 128.7134s and 3.9499s respectively. The calculated time for this recognition system includes: (1) Apex frame spotting using *divide-and-conquer* strategy; (2) Horizontal and vertical components of optical flow estimation; (3) Orientation, magnitude and optical strain images computation; (4) Bi-WOOF histogram generation; (5) Expression classification in SVM. Both of the experiments were carried out on an Intel Core i7-4770 CPU 3.40GHz processor. For the case of *two images*, it achieved a speedup of approximately 97%, or in other words, ~ 33 times faster compared to the *whole sequence*. It is indisputable that extracting the features from only *two images* is faster than the *whole sequence* due to lesser images involved in the system, and hence reduce the complexity.

5 Conclusion

In the recent few years, a number of research groups have attempted to improve the accuracy of micro-expression recognition by designing a variety of feature extractors that can best capture the subtle facial changes [29, 12, 21], while a few other works [13, 15, 17] have sought out ways to reduce information redundancy in micro-expressions (using only a portion of all frames) before recognizing them.

In this paper, we demonstrated that it is sufficient to encode facial micro-expression features by utilizing only the apex frame (and onset frame as reference frame) from among all frames of an entire video sequence. To the best of our knowledge, this is the first attempt at recognizing micro-expressions from video using only the apex frame. For databases that do not provide apex frame annotations, the apex frame can be acquired by automatic spotting based on a *divide-and-conquer* search strategy employed in our recent work [20]. We also proposed a novel feature extractor, Bi-Weighted Oriented Optical Flow (Bi-WOOF), which can concisely describe discriminately weighted motion features extracted from the apex and onset frames. As its name implies, the optical flow histogram features (bins) are locally weighted by their own magnitudes while facial regions (blocks) are globally weighted by the magnitude of optical strain, a reliable measure of subtle deformation.

Experiments conducted on four publicly available micro-expression databases—CASME II, SMIC-HS, SMIC-NIR and SMIC-VIS, demonstrated the effectiveness and efficiency of the proposed approach. Using a single apex frame for micro-expression recognition, the two high frame rate databases, CASME II and SMIC-HS both achieved the highest recognition rate of 61% and 62% respectively, compared to recent state-of-the-art methods reported in literature.

5.1 *Prima facie*

In this work, we have established two strong propositions, which are by no means conclusive at this juncture as further research is necessary:

1. **The apex frame is the most important frame in a micro-expression clip**, that it contains the most intense or expressive micro-expression information. Our experiments by random frame selection (as the supposed apex frame) substantiates this fact. Perhaps, it will be interesting to know to what extent an imprecise apex frame (for example, a detected apex frame that is located a few frames away) could influence the recognition performance.
2. **The apex frame is sufficient for micro-expression recognition.** A majority of recent state-of-the-art methods promote the use of the entire video sequence, or a reduced set of frames [17, 15]. In this work, we advocate the opposite, that "less is more", backed by our hypothesis that a large number of frames is not necessary to guarantee a high recognition accuracy, particularly in the case when high-speed cameras are employed (for CASME II and SMIC-HS). Comparisons against conventional methods show that the use of a well-spotted apex frame can provide better information than an array of frames. At this juncture, it is premature to ascertain the reasons behind this finding. Hence, this warrants a detailed investigation into *how* and *where* micro-expression cues reside within the sequence itself.

References

1. Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition*, pages 3444–3451, June 2013.
2. Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.
3. Dalal, N., Triggs, B., and Schmid, C. Human detection using oriented histograms of flow and appearance. In *Proc. of ECCV*, pages 428–441, 2006.
4. Davison, A. K., Yap, M. H., and Lansley, C. Micro-facial movement detection using individualised baselines and histogram-based descriptors. In *Systems, Man, and Cybernetics (SMC)*, pages 1864–1869, October 2015.
5. Ekman, P. Lie catching and microexpressions. *The philosophy of deception*, pages 118–133, 2009.
6. Ekman, P. and Friesen, W. V. Nonverbal leakage and clues to deception. *Journal for the Study of Interpersonal Processes*, 32:88–106, 1969.
7. Ekman, P. and Friesen, W. V. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.
8. Ekman, P. and Friesen, W. V. *Facial action coding system*. Consulting Psychologists Press, 1978.
9. Frank, M. G., Herbasz, M., Sinuk, K., Keller, A., Kurylo, A., and Nolan, C. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In *Annual meeting of the International Communication Association, Sheraton New York, New York City, NY*, 2009.
10. Frank, M. G., Maccario, C. J., and Govindaraju, V. *Protecting Airline Passengers in the Age of Terrorism*, pages 86–106. ABC-CLIO, 2009.
11. Goshtasby, A. Image registration by local approximation methods. *Image and Vision Computing*, 6(4):255–261, 1988.
12. Huang, X., Wang, S. J., Zhao, G., and Pietikainen, M. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In *ICCV Workshops*, pages 1–9, 2015.
13. Le Ngo, A. C., Liong, S. T., See, J., and Phan, R. C. W. Are subtle expressions too sparse to recognize? In *Digital Signal Processing (DSP)*, pages 1246–1250, July 2015.
14. Le Ngo, A. C., Phan, R. C. W., and See, J. Spontaneous subtle expression recognition: Imbalanced databases & solutions. In *Asian Conference on Computer Vision*, pages 33–48, 2014.
15. Le Ngo, A. C., See, J., and Phan, R. C. W. Sparsity in dynamics of spontaneous subtle emotions: Analysis & application. *IEEE Transactions on Affective Computing*, 2016.
16. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., and Pietikinen, M. Reading hidden emotions: Spontaneous micro-expression spotting and recognition. *arXiv preprint arXiv:1511.00423*, 2015.
17. Li, X., Pfister, T., Huang, X., Zhao, G., and Pietikainen, M. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition*, pages 1–6, April 2013.
18. Liong, S. T., Phan, R. C.-W., See, J., Oh, Y. H., and Wong, K. Optical strain based recognition of subtle emotions. In *International Symposium on Intelligent Signal Processing and Communication Systems*, pages 180–184, December 2014.
19. Liong, S. T., See, J., Phan, R. C. W., Le Ngo, A. C., Oh, Y. H., and Wong, K. Subtle expression recognition using optical strain weighted features. In *Asian Conference on Computer Vision Workshops on Computer Vision for Affective Computing*, pages 644–657, November 2014.
20. Liong, S. T., See, J., Wong, K., A.C. Le Ngo, A. C., Oh, Y. H., and Phan, R. C. W. Automatic apex frame spotting in micro-expression database. In *Asian Conference on Pattern Recognition (ACPR)*, 2015.
21. Liu, Y. J., Zhang, J. K., Yan, W. J., Wang, S. J., Zhao, G., and Fu, X. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transaction of Affective Computing*, 2016.

22. Moilanen, A., Zhao, G., and Pietikainen, M. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *International Conference on Pattern Recognition (ICPR)*, pages 1722–1727, August 2014.
23. Oh, Y. H., Le Ngo, A. C., See, J., Liong, S. T., Phan, R. C. W., and Ling, H. C. Monogenic riesz wavelet representation for micro-expression recognition. In *Digital Signal Processing*, pages 1237–1241. IEEE, July 2015.
24. O’Sullivan, M., Frank, M. G., Hurley, C. M., and Tiwana, J. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):530–538, 2009.
25. Porter, S. and ten Brinke, L. Reading between the lies identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, 19(5):508–514, 2008.
26. Shreve, M., Brizzi, J., Fefilatyev, S., Luguev, T., Goldgof, D., and Sarkar, S. Automatic expression spotting in videos. *Image and Vision Computing*, 32(8):476–486, 2014.
27. Shreve, M., Godavarthy, S., Manohar, V., Goldgof, D., and Sarkar, S. Towards macro-and micro-expression spotting in video using strain patterns. In *Applications of Computer Vision (WACV)*, pages 1–6, 2009.
28. Simof, J. C. and Hughes, T. J. R. *Computational Inelasticity*. Springer, 2008.
29. Wang, Y., See, J., Phan, R. C. W., and Oh, Y. H. LBP with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Computer Vision—ACCV*, pages 525–537, 2014.
30. Xu, F., Zhang, J., and Wang, J. Microexpression identification and categorization using a facial dynamics map. *IEEE Transactions on Affective Computing*, 2016.
31. Yamaji, A. *An Introduction to Tectonophysics: Theoretical Aspects of Structural Geology*. Terra-pub, 2007.
32. Yan, W. J., Wang, S. J., Chen, Y. H., Zhao, G., and Fu, X. Quantifying micro-expressions with constraint local model and local binary pattern. In *Computer Vision-ECCV workshop*, pages 296–305, September 2014.
33. Yan, W.-J., Wang, S.-J., Zhao, G., Li, X., Liu, Y.-J., Chen, Y.-H., and Fu, X. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE*, 9:e86041, 2014.
34. Zach, C., Pock, T., and Bischof, H. A duality based approach for realtime TV-L1 optical flow. pages 214–223, 2007.